US010585989B1

(12) **United States Patent**
Ahmed et al.

(10) **Patent No.: US 10,585,989 B1**
(45) **Date of Patent: Mar. 10, 2020**

(54) **MACHINE-LEARNING BASED DETECTION AND CLASSIFICATION OF PERSONALLY IDENTIFIABLE INFORMATION**

(71) Applicant: **International Business Machines Corporation**, Armonk, NY (US)

(72) Inventors: **Mohamed N. Ahmed**, Loudoun County, VA (US); **Andeep S. Toor**, Chantilly, VA (US)

(73) Assignee: **INTERNATIONAL BUSINESS MACHINES CORPORATION**, Armonk, NY (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **16/125,389**

(22) Filed: **Sep. 7, 2018**

(51) **Int. Cl.**
G06F 17/27 (2006.01)
G06N 3/04 (2006.01)
G06N 3/08 (2006.01)

(52) **U.S. Cl.**
CPC ***G06F 17/2785*** (2013.01); ***G06F 17/274*** (2013.01); ***G06F 17/2735*** (2013.01); ***G06N 3/0445*** (2013.01); ***G06N 3/0454*** (2013.01); ***G06N 3/08*** (2013.01)

(58) **Field of Classification Search**
CPC G06F 17/2735; G06F 17/274; G06F 17/2785; G06N 3/08; G06N 3/0445; G06N 3/0454
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

8,234,263	B2	7/2012	Pradhan et al.	
9,396,179	B2	7/2016	Stavrianou et al.	
9,805,371	B1 *	10/2017	Sapoznik	G06F 16/9024
10,332,508	B1 *	6/2019	Hoffmeister	G10L 15/16
2008/0263333	A1	10/2008	Wang et al.	
2014/0068706	A1 *	3/2014	Aissi	G06F 21/6254 726/1
2014/0074845	A1	3/2014	Dimassimo et al.	
2014/0164408	A1	6/2014	Dubbels	
2017/0316285	A1 *	11/2017	Ahmed	G06K 9/66
2018/0075254	A1 *	3/2018	Reid	G06F 7/00
2019/0018983	A1 *	1/2019	Anderson	G06F 21/64
2019/0080063	A1 *	3/2019	Rice	G06F 21/316
2019/0171846	A1 *	6/2019	Conikee	G06F 21/6245

* cited by examiner

Primary Examiner — Walter Yehl
(74) *Attorney, Agent, or Firm* — Garg Law Firm, PLLC; Rakesh Garg; James Nock

(57) **ABSTRACT**

Detection and classification of personally identifiable information includes identifying a document with a known author. A first set of features of the document is extracted using natural language processing, and a second set of features of the document is extracted based upon one or more past documents for the known author using a recurrent neural network. The first set of features and the second set of features are classified using a classifier to produce classified extracted features. Personally identifiable information is labeled in the document based upon the classified extracted features.

19 Claims, 7 Drawing Sheets